zBoston Transportation Analysis: Fastest Route to Northeastern University

Charlie Deane **Peter Li**

Abstract

Regression analysis on the arrival times and headways of Boston's public transportation system using MBTA and local weather data. Modeling and predicting arrival times and headways using regression models and DNNs.

Introduction

Historically the authors of this paper have been unpunctual when attending their Data Mining & Machine Learning course. While this is in part due to their lackluster sleep schedule, some blame can be put on the irregularity of the Boston transportation system. With this motivating reason, we outline a proposal to analyze Boston's subway and bus system to maximize the likelihood of the authors' timely arrival to class.

There exist many modern tools to estimate travel times like Google Maps, Apple Maps, and the MBTA API, which are *generally* accurate for roadway congestion. However, these applications can be especially unreliable when estimating the public transit arrivals and headways. Public transportation is vital in densely populated cities like Boston and is especially important for the faculty and student population at Northeastern. Therefore, our analysis could be potentially abstracted to provide the residents of Boston an accurate estimation of public transportation arrival times and headways.

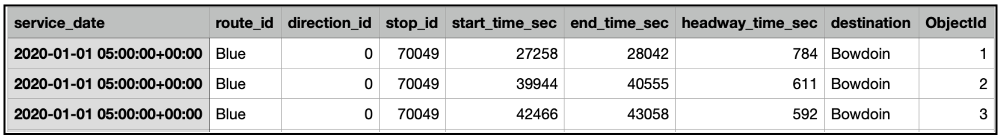
While we would like to model the entire transportation system of Boston, we realize that such a project would be beyond our scope and resources. Therefore we modify our starting location to be the neighboring residential areas of Northeastern University (e.g. Mission Hill, Fenway, Symphony). Our model and analysis will focus on the route from these neighborhoods to public transportation stops near Northeastern University.

Proposed Project

We will utilize regression to analyze the efficiency of different transportation methods to the Northeastern campus. To begin, we plan to utilize a variety of transportation data pertaining to train networks, bus routes, and their respective arrival and headway times provided by the Massachusetts Bay Transportation Authority (MBTA). These datasets will provide us with the required headways, arrival times, and locations. We realize that weather can influence the punctuality of public transportation; therefore, we plan to include historical weather data (precipitation, highs, lows) for each day of the calendar year.

Training Our Model

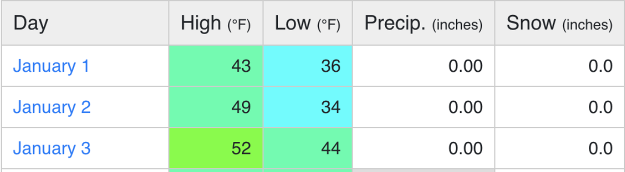
Since we are constrained by computational resources we will limit our training data to 800 randomly selected samples (time points) and our test set to 200 samples (time points). Our feature vector will include information pertaining to: the time of day, location, and weather data (for a minimum of 5 features). Our features are mostly extracted (see Figure 1.1) from the provided datasets; however we will have to do some preprocessing of our information before using them in our model. First, the given times are in second, not the optimal format for our regression model. So we will have to convert the times to minutes, or to a standard date time format.



service_date	route_id	direction_id	stop_id	start_time_sec	end_time_sec	headway_time_sec	destination	ObjectId
2020-01-01 05:00:00+00:00	Blue	0	70049	27258	28042	784	Bowdoin	1
2020-01-01 05:00:00+00:00	Blue	0	70049	39944	40555	611	Bowdoin	2
2020-01-01 05:00:00+00:00	Blue	0	70049	42466	43058	592	Bowdoin	3

Figure 1.1 Rapid-Transit Headways

Second, our weather dataset (ExtremeWeatherWatch) does not provide a readable CSV file, so we will have to do some copy pasting into our own CSV file. Finally, the weather dataset provided does not give the exact weather conditions at the specific arrival and departure times from the MBTA dataset, so we will have to extrapolate the entire day's weather information to that specific time point.



Day	High (°F)	Low (°F)	Precip. (inches)	Snow (inches)
January 1	43	36	0.00	0.0
January 2	49	34	0.00	0.0
January 3	52	44	0.00	0.0

Figure 1.2 Weather Dataset

If we have additional time, we would like to find the actual time it takes to get to Northeastern. This would incorporate calculating the actual path from each location to Northeastern University (this is not provided to us) and adjusting for line changes, layovers, and intermediate travel distances.

Regression models we plan to include: vanilla linear regression, robust regression/ridge regression, logistic regression, DNNs.

Dataset:
We chose the year 2020, because it was one of the only complete datasets available for headways and arrivals for the Rapid-Transit and Bus.

Rapid-Transit Headways 2020
Headers: route_id, direction_id, start_time_sec, end_time_sec, headway_time_sec, destination, ObjectId
Samples: Over 400 samples for each transit stop every day of the calendar year
https://mbta-massdot.opendata.arcgis.com/datasets/mbta-rapid-transit-headways-2020-1/about

MBTA Bus Arrival and Departures 2020
Headers: service_date, route_id, direction_id, half_trip_id, stop_id, time_point_id, time_point_order, point_type, standard_type, scheduled, scheduled_headway, headway
Samples: Over 2000 samples for each route every day of the calendar year
https://mbta-massdot.opendata.arcgis.com/datasets/mbta-bus-arrival-departure-times-2020/about

Rapid-Transit Arrival and Departures 2020
Headers: service_date, route_id, trip_id, direction_id, stop_id, stop_sequence, vehicle_id, vehicle_label, event_type, event_time, event_time_sec
Samples: Over 400 samples for each transit stop every day of the calendar year
https://mbta-massdot.opendata.arcgis.com/datasets/mbta-rapid-transit-events-2020/about

Boston Regional Weather 2020
Headers: High (F), Low (F), Precipitation (in), Snow (in)
Samples: 1 sample each day of the calendar year for a total of 365 samples
https://www.extremeweatherwatch.com/cities/boston/year-2020

Project Timeline

10/28/2022	Submit project proposal
11/04/2022	Evaluate proposal feedback and decide to either move forward or submit a new proposal by 11/08/2022.
11/11/2022	Create a shared CoLab Notebook and import the data set using SciKit and Pytorch
11/15/2022	Begin to train models, and over course the week, tweak any changes that may need to be made
11/30/2022	Finish training on the dataset, and begin to work on the post analysis of the training output.
12/06/2022	Finish analysis of data, and begin to work on review of the methodology, alongside the investigation of the performance comparison against different methods we defined prior
12/13/2022	Meet in person one last time and submit the project.